Investigation comparing simulation methods for models with common constraints

Sean van der Merwe

University of the Free State

28 November 2018

Statistics are not the point of statistics and that is the main point of my talk*

*Warning: this talk may contain traces of statistics

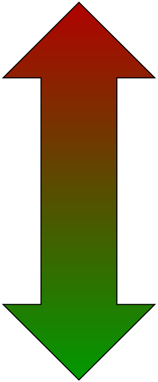
seanvdm.co.za

UNIVERSITY OF THE FREE STATE
UNIVERSITEIT VAN DIE VRYSTAAT
YUNIVESITHI YA FREISTATA
UFS UV

Overview

A scale of deceit

Most deceitful



Least deceitful

- Politicians on the other side
- Politicians on your side
- People trying to sell you something
- Regular people
- Bad statisticians
- Bad data
- Good data
- Good statisticians
- Good statisticians using good models

seanvdm.co.za

What makes a good statistician?

- My opinion: talking/writing **less nonsense** than average
- By doing the very thing that separates statistics from other fields: **quantifying uncertainty**
- Building and fitting models is not unique to statistics
  - Every field has specialist knowledge
  - But to them the errors are a nuisance
- For a statistician the uncertainty is what matters most
- Because uncertainty stops us from saying things are meaningful when they could be chance/noise/never to happen again
- Can also give us confidence to say things are meaningful based on uncertain evidence

seanvdm.co.za

Why uncertainty matters

- It may be a cliché, but a point estimate is always wrong
- Consider the USA election in 2016:
  - First people trusted polls so much they didn't vote
  - Then they got angry that the results were not as polled
  - In truth the results were well within the estimated uncertainty
- Example from French and Garielli (2005):
  - If you're buying or selling a house then you see 1 number, loosely based on somebody's idea of the value of the house
  - Sometimes it's highly negotiable, sometimes it's not.
  - Sometimes it's a conservative estimate, sometimes it's an aggressive estimate
  - How do we communicate that?

Quantifying uncertainty

To quantify uncertainty we must consider **ALL** the uncertainty

- Measurement error
- Non-random sampling
- Missing information
- Model misspecification
- Parameter estimation
- Residual error
- Systematic changes over time

Cressie *et al.* (2009) suggest that the hierarchical statistical model is the way forward, and I agree.

Fitting these models

- It is possible to construct models that use all appropriate uncertainty
- The easiest way is the **BUGS** framework
- Not necessarily the best but definitely the most flexible
- The notation is hierarchical so it's perfect for hierarchical models
    1. Specify distribution for observations
    2. Specify parameters of distribution (can be in terms of explanatory variables)
    3. Specify priors for parameters not defined in terms of other parameters
    4. Supply data and hyperparameters
    5. Supply initial values
    6. Let BUGS do the work
    7. Analyse results
    8. Answer questions

seanvdm.co.za

Overview

Identifiable parameters

- With flexibility come options and choices (can be overwhelming)
- Error messages are not always descriptive
- New problems come up not previously considered
- Luo *et al.* (2009) go on to discuss the issues of identifiability and equifinality in detail
- Rannala (2002) explains that Bayesian models are theoretically identifiable as long as the posterior is proper, but what about the practical sense?
- The problem can be summed up (in my words) as:

If different models or different parameter values fit data equally well then which values do we assign to those parameters and what do those parameters then mean?

Overview

First example

- Consider $Y_i \sim N\left(\alpha_{g_i} + \beta_{h_i}, \sigma^2\right)$
  - g_i indicates the level of the first factor to which observation i belongs
  - h_i is the level of the second factor to which observation i belongs
  - α values and β values are thus fixed effects to be estimated, one for each level of each factor
- This is closest to the way in which the model is programmed for maximum flexibility
- We can add a constant to each α and then subtract that same constant from each β without affecting any expected values
- Thus, the interpretation of any specific α value is not clear

The solutions

1. Set an arbitrary fixed effect to zero, *e.g.* $\alpha_1 = 0$.
2. Force a set of fixed effects to sum to zero by setting one equal to negative the sum of the others, *e.g.* $\alpha_{n_g} = -\sum_{g=0}^{n_g-1} \alpha_g$
3. Pull a set of fixed effects to sum to zero, *e.g.* $\sum_g \alpha_g \sim N(0, 1/\tau)$

Solution 3 is the new approach I'm introducing, in the hope of improving both fit and simulation speed for Bayesian implementations of the model

seanvdm.co.za

Effect

- The obvious question: does it make a difference?
- In this simple case: no
- It doesn't hurt either
- But what if we make it more complicated?
- If we replace the Normal errors with Student-t errors:
  - All solutions have good accuracy
  - All solutions have similar simulation time
  - Using BUGS does provide an advantage though — changing from Normal to t errors in the model is a 1 to 2 minute process to change the code

Second example

- Let's take it up another notch
- What if we have multivariate data with constraints across dimensions?
- With compositional data you have vectors that must sum to one, with all components positive
- The standard approach is to transform the data using one component as a reference component (see Maier, 2014)
- But what if you don't have a natural reference component?
- I'm suggesting using the Bayesian hierarchical model approach with a penalty instead
- You can analyse any and all dimensions directly
- You can answer any sort of question without making additional assumptions
- You even get a better fit

seanvdm.co.za

My approach

- Results from two simulated scenarios are given, the first is a MANOVA problem, while the second also includes regressors

Scenario A	Target	Old approach	New approach
Error	0	19.59	18.38
Coverage	0.95	0.87	0.94
Scenario B	Target	Old approach	New approach
Error	0	19.19	18.81
Coverage	0.95	0.85	0.86

- Even more impressive is the gain in inferential power:
  - For the regressors in Scenario B, the median p-values for the existing method are 50% and 24%, while the new method yields 1%, 0.1%, and 0.4%
- Thus, my approach lets me test all dimensions and correctly picks up relationships

UNIVERSITY OF THE FREE STATE UNIVERSITEIT VAN DIE VRYSTAAT YUNIVESITHI YA FREISTATA UFS UV

seanvdm.co.za

Overview

The point

- Calculating and comparing numbers is not the point of statistics
- The point is to draw appropriate information from numbers in order to answer questions that matter
- To give good answers you must understand **and convey** the uncertainty in your information
- Bayes does that best
- Bayes can be easy with appropriate use of modern tools

For more information on this work, or to download this presentation, visit my website at seanvdm.co.za/project/thesisstuff/

Thank you for listening!

seanvdm.co.za

UNIVERSITY OF THE FREE STATE
UNIVERSITEIT VAN DIE VRYSTAAT
YUNIVESITHI YA FREISTATA
UFS UV